21-49-004 #7

UNITED STATES AIR FORCE

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

Reproduced From
Best Available Copy

School of AVIATION MEDICINE

DISCRIMINATORY ANALYSIS
VII. On the Theory of Selection

PROJECT NUMBER 21-49-004
REPORT NUMBER 7

19990714 143

PROJECT REPORT

DISCRIMINATORY ANALYSIS
VII. On the Theory of Selection

E.L. LEHMANN, Ph.D.

University of California, Berkeley

PROJECT NUMBER 21-49-004
REPORT NUMBER 7

USAF SCHOOL OF AVIATION MEDICINE
RANDOLPH FIELD, TEXAS
AUGUST 1951

ON THE THEORY OF SELECTION

1. Introduction.

The present paper is the third of a series of papers on the classification problem. The first paper is concerned with the classification of individuals. The second paper, not yet completed, will deal with the classification of populations.

In this third paper we shall consider a special class of classification problems with no standards given. Again each of s populations $\pi_1, \cdots, \pi_s$ is to be classified into one of two categories ("good" and "bad"). However, what constitutes a good or bad population is not defined absolutely but in terms of the quality of the populations at hand.

Problems of this kind arise frequently and have recently been treated by a number of authors (Mosteller [1], Paulson [2], Stein [3], Bahadur [4]). We assume that we have a sample $X_{ij} (j = 1, \cdots, n_i)$ from each of the populations π_i and that the distribution of the X_{ij} depends on an unknown parameter θ_i. The populations are ranked according to the values of the θ's. (For example, if the θ's are real valued, that one of two populations may be the better that has the higher θ-value.)

This work was done at Columbia University and supported in part by the USAF School of Aviation Medicine. The first paper of this series, referred to above, is Report No. 6 of Project No. 21-49-004 ("On the Simultaneous Classification of Several Individuals"), USAF School of Aviation Medicine.

As has been pointed out by Bahadur, it is frequently desirable to go beyond classifying the populations as good or bad. If, for example, one is looking for good varieties of wheat to plant, one must decide in what proportions to plant them. However, this more detailed analysis is not always required. When it does apply, the solution tends to be that procedure which is appropriate when, rather than selecting a number of good populations, one is interested only in the best of them.

We shall here follow essentially the formulation of Paulson who considered this problem for normal populations. To be specific, let us assume that the θ's are real valued and that quality improves with increasing θ. Let us assume further that there is given a function $g(\theta, \theta')$ increasing in the second variable and decreasing in the first, such that the population π_i is considered good provided

$$g(\theta_i, \max_j \theta_j) < \Delta$$

where Δ is some fixed positive number. If the variables are normally distributed with mean θ and variance σ^2, we may for example take $g(\theta, \theta') = \theta' - \theta$ or $g(\theta, \theta') = \frac{\theta' - \theta}{\sigma}$. In the Poisson case we might take $g(\theta, \theta') = \theta'/\theta$ and in binomial case

$$g(\theta, \theta') = \frac{\frac{\theta}{1-\theta}}{\frac{\theta'}{1-\theta'}},$$

where in both cases θ indicates the mean of the variables.

As before we shall adopt the point of view of the Neyman-Pearson theory and ask that

(1.1) the expected number of bad populations classified as good $\overset{be}{\underset{\wedge}{\leq}} \gamma$.

Subject to (1.1) we wish to maximize the expected number of good populations classified correctly.

Condition (1.1) has one consequence that may seem undesirable. It follows from the definition just given, that there is always at least one good population, that with the maximum θ. However, if we impose (1.1) we may sometimes have to classify all of the populations as bad. This will occur, roughly speaking, when the observations indicate a situation in which the sample size is too small to make the selection of the good populations with the desired degree of accuracy. There are two ways of avoiding this difficulty. If one knows the order of magnitude of the parameters involved, one can determine a sample size which makes it very probable that one will be able to perform the classification. Alternatively, of course, the situation points to the use of sequential procedures. Such procedures also have the advantage for problems of the kind considered here that they permit classifying the populations gradually. A decision will be reached early on those that are either very good or very poor, while for the

intermediate ones one may take a larger number of observations. A procedure of this kind was discussed by Stein [3].

Although it is easy to develop a general theory of minimax procedures for the classification problem described here, the application of such a theory to particular cases runs into difficulties which the author so far has not been able to surmount. We shall illustrate the situation with an example.

2. An example.

Perhaps the simplest example, and one which was considered by Paulson, assumes that the X_{ij}, $i=1,\cdots,s$; $j=1,\cdots,n$ are samples from normal distributions with means θ_i and common variance $\sigma^2 = 1$. Since the $\bar{X}_i = \frac{1}{n}\sum_j X_{ij}$ form a set of sufficient statistics we assume without loss of generality that $n = 1$ and denote our variables by $X_1, \cdots, X_s$. We take $g(\theta, \theta') = \theta' - \theta$, so that π_i is considered good if $\theta_i > \max_j \theta_j - \Delta$.

In order to obtain the minimax procedure we must guess two least favorable distributions: one that maximizes on the average the number of good populations that are classified as bad, and one that maximizes the number of bad populations that are classified as good. One conjectures that both are concentrated on the set

$$\begin{aligned} &s - 1 \text{ of the } \theta\text{'s are equal, say} = \theta \\ &\text{the remaining mean} = \theta + \Delta \ . \end{aligned}$$

It turns out that it is immaterial what value of θ we take. Let us put $\theta = 0$, and let us assume that the a priori distribution assigns probability $\frac{1}{s}$ to each of the possible parameter sets $(\Delta, 0, \cdots, 0), (0, \Delta, 0, \cdots, 0), \cdots (0, \cdots, 0, \Delta)$. Then the Bayes problem becomes

Maximize

$$(1.2) \qquad \frac{1}{s} \sum_i E\left[\phi_1(X) + \cdots + \phi_s(X) \mid \theta_i = \Delta, \ \theta_j = 0 \text{ if } j \neq i\right]$$

subject to the condition

$$(1.3) \qquad \frac{1}{s} \sum_i E\left[\sum_{j=1}^{s} \phi_j(X) - \phi_i(X) \mid \theta_i = \Delta, \ \theta_j = 0 \text{ if } j \neq i\right] \leq \gamma$$

This problem is solved by means of a lemma to be proved in the paper on the classification of populations. The Bayes solution sets $\phi_1(x) = 1$ if

$$e^{-\frac{1}{2}(x_1-\Delta)^2 - \frac{1}{2}(x_2^2+\cdots+x_s^2)} + e^{-\frac{1}{2}(x_2-\Delta)^2 - \frac{1}{2}(x_1^2+x_3^2+\cdots+x_s^2)} + \cdots$$

$$> k\left[e^{-\frac{1}{2}(x_2-\Delta)^2 - \frac{1}{2}(x_1^2+\cdots+x_s^2)} + e^{-\frac{1}{2}(x_3-\Delta)^2 - \frac{1}{2}(x_1^2+x_2^2+x_4^2+\cdots+x_s^2)} + \cdots\right]$$

that is, if

$$e^{\Delta x_1} + e^{\Delta x_2} + \cdots + e^{\Delta x_s} > k'\left[e^{\Delta x_2} + \cdots + e^{\Delta x_s}\right]$$

and hence if

$$e^{\Delta X_2} + \cdots + e^{\Delta X_s} < C\, e^{\Delta X_1}.$$

The solution for the remaining ϕ's is obtained by symmetry.

To complete the solution we must now show that for this procedure

(a) the expected number of good populations classified as bad takes on its maximum when $\theta_s = \Delta$, $\theta_1 = \cdots = \theta_{s-1} = 0_+$

(b) the expected number of poor populations classified as good takes on its maximum when $\theta_s = \Delta$, $\theta_1 = \cdots = \theta_{s-1} = 0_-$.

We need to prove (b) not only to show that the procedure is minimax but also to show that the correct determination of C is

$$(1.4)\quad P(e^{\Delta X_2} + \cdots + e^{\Delta X_s} < C\, e^{\Delta X_1} \mid \theta_1 = \cdots = \theta_{s-1} = \theta_s - \Delta = 0) = \frac{k}{s} = \alpha$$

The difficulty referred to in the beginning of the last section consists in the proof of (a) and (b). We shall now present certain partial results on this problem.

Let us first consider what happens for large Δ. Let us set $C = e^{\gamma}$ and determine it by means of (1.4). If we put $Y_i = X_i - \theta_i$, (1.4) becomes

$$P(e^{\Delta Y_2} + \cdots + e^{\Delta Y_{s-1}} + e^{\Delta(Y_s+\Delta)} < e^{\Delta Y_1 + \gamma}) = \alpha$$

Dividing both sides by e^{Δ^2}, we see that for large Δ this becomes essentially

$$P(e^{\Delta Y_s} < e^{\Delta Y_1 + \gamma - \Delta^2}) = \alpha$$

Therefore $\dfrac{\gamma - \Delta^2}{\Delta} \to k \quad (k>0)$ as $\Delta \to \infty$.

We shall first consider problem (a). As $\Delta \to \infty$, the expected number of good populations classified as good when $\theta_1 = \cdots = \theta_{s-1} = 0_-$, $\theta_s = \Delta$ tends to 1.

It is clear that if $s-1$ of the populations are bad, and only one good, the probability of classifying the good one correctly is minimized when $\theta_1 = \cdots = \theta_{s-1} = 0$; i.e. when we are in the situation which we believe to be the minimizing one.

We need to compare this situation with those in which there are more than one good population Suppose $\theta_s = \Delta$, $\theta_1 \geq 0_+$, $\theta_2, \cdots, \theta_{s-1} \leq \Delta$.

Let

$$P_1 = P(e^{\Delta(Y_2+\theta_2)} + \cdots + e^{\Delta(Y_s+\theta_s)} < e^{\Delta(Y_1+\theta_1)} + \gamma$$

and let $P_2, \cdots, P_s$ be defined analogously. The expectation we are concerned with is the sum of those P's corresponding to good populations and hence is $\geq P_1 + P_s$.

We shall now show that uniformly for

$\theta_1 \geq 0_+, \theta_2, \cdots, \theta_{s-1} \leq \Delta$ we have

$$P_s \to 1$$

$$P_1 \geq M_1(\Delta)$$

where $\lim_{\Delta \to \infty} M_1(\Delta) > 0$.

$$P_s = P(e^{\Delta(Y_1+\theta_1)} + \cdots + e^{\Delta(Y_{s-1}+\theta_{s-1})} < e^{\Delta(Y_s+\theta_s)+\gamma})$$

$$\geq P(e^{\Delta Y_1} + \cdots + e^{\Delta Y_{s-1}} < e^{\Delta Y_s + \gamma(\Delta)}) \to \infty$$

since $\gamma = \Delta^2 - k(\Delta) + o(\Delta)$.

On the other hand

$$P_1 > P(e^{\Delta(Y_2+\Delta)} + \cdots + e^{\Delta(Y_s+\Delta)} < e^{\Delta Y_1+\Delta^2-k\Delta+o(\Delta)})$$

$$= P(e^{\Delta Y_2} + \cdots + e^{\Delta Y_s} < e^{\Delta Y_1-k\Delta+o(\Delta)})$$

$$\frac{e^{\Delta Y_2}+\cdots+e^{\Delta Y_s}}{s-1} < e^{\Delta\left(\frac{Y_2+\cdots+Y_s}{s-1}\right)}$$

$$\therefore P_1 > P\left((s-1)\, e^{\Delta\frac{Y_2+\cdots+Y_s}{s-1}} < e^{\Delta Y_1-k\Delta+o(\Delta)}\right)$$

$$= P\left(\frac{Y_2+\cdots+Y_s}{s-1} + \frac{\log(s-1)}{\Delta} < Y_1 - k + \frac{o(\Delta)}{\Delta}\right)$$

$$\rightarrow P\left(\frac{Y_2+\cdots+Y_s}{s-1} < Y_1 - k\right) > 0.$$

Unfortunately we have not been able to show as much for the even more important problem (b), and in fact it seems doubtful that the result (b) holds exactly, even for large Δ What we can show here is that the value $\hat{\theta}_1, \cdots, \hat{\theta}_{s-1}$ that maximize the expected number of bad populations classified as good are such that for $i = 1, \cdots, s-1$

$$\hat{\theta}_i \rightarrow 0_- \quad \text{as } \Delta \rightarrow \infty .$$

The proof of this is in fact quite simple. It is obvious that $\hat{\theta}_1, \cdots, \hat{\theta}_{s-1}$ are all ≤ 0. The expectation we are concerned with is thus $P_1+\cdots+P_{s-1}$.

Consider now a sequence of values $\theta_1^{(\Delta)}$ tending to a limiting value θ_1. Then it is seen that regardless of the values of $\theta_2, \cdots, \theta_{s-1} \leq 0$ (but not necessarily uniformly in these variables)

$$\lim_{\Delta \rightarrow \infty} P_1 = P(Y_s < Y_1 + \theta_1 - k)$$

Thus in the limit P_1 is maximized for $\theta_1 = 0$ regardless of the other θ's, and similarly for $\theta_2, \cdots, \theta_{s-1}$. Hence the result follows.

This proves that at least asymptotically the procedure has the correct size.

References

[1] F. Mosteller, "A k-sample slippage test for an extreme population," Annals of Math. Stat., Vol. 19 (1948), p. 58.

[2] E. Paulson, "A multiple decision procedure for certain problems in the analysis of variance," Annals of Math. Stat., Vol. 20 (1949), p. 95.

[3] C. Stein, "On the selection of the largest of a set of means," Annals of Math. Stat., Vol. 19 (1948), p. 429.

[4] R. Bahadur, "On a problem in the theory of k populations," Annals of Math. Stat., Vol. 21 (1950), p. 362.

Distributed by
USAF School of Aviation Medicine
Distribution Number 1680
August 1951

DEFENSE—SAN ANTONIO